Bidirectional Cross-Modal Knowledge Exploration for Video Recognition with Pre-trained Vision-Language Models

Wenhao Wu[1,2] *Xiaohan Wang*[3] *Haipeng Luo*[4] *Jingdong Wang*[2] *Yi Yang*[3] *Wanli Ouyang*[5,1]

[1]The University of Sydney [2]Baidu Inc. [3]Zhejiang University [4]UCAS [5]Shanghai AI Laboratory

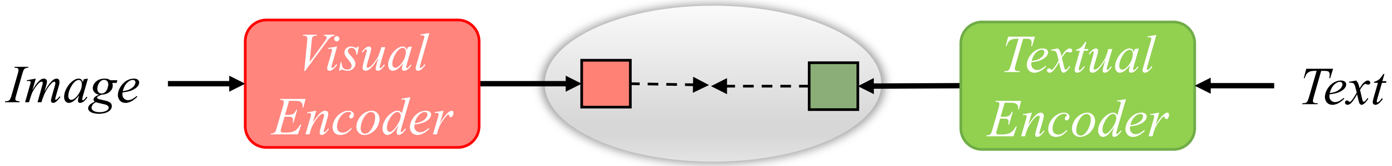
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Code & Models

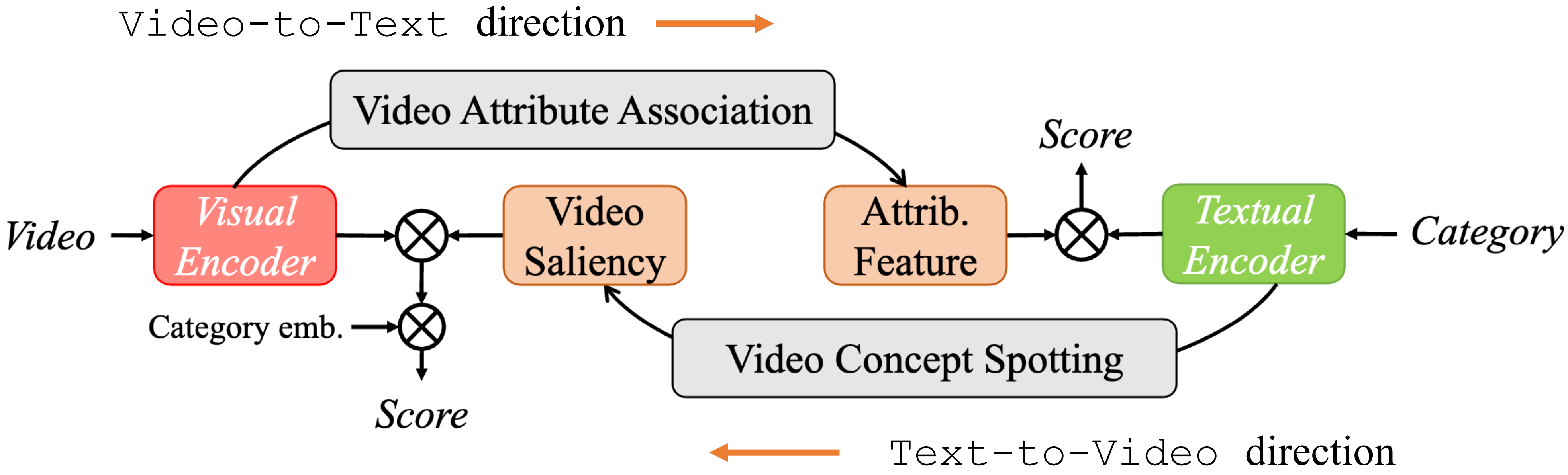
Poster : TUE-PM-238

Key Innovation

(a) Pre-trained Vision-Language Models (VLMs) build a bridge between the visual and textual domains.

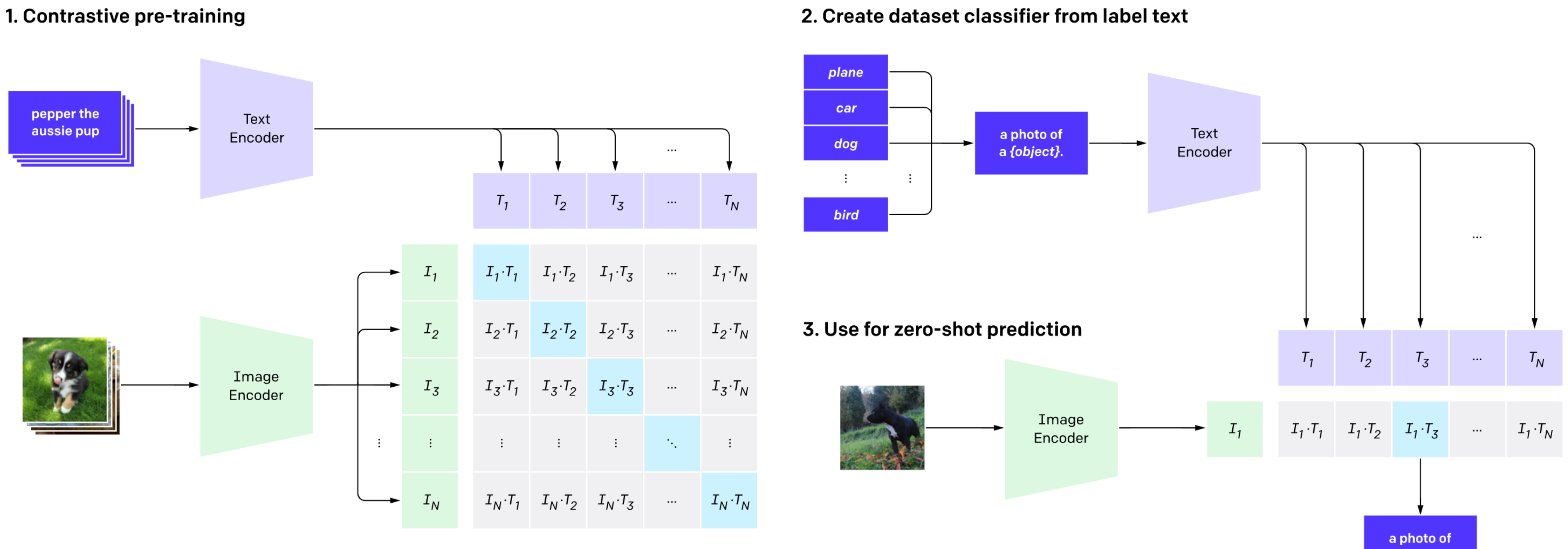


(b) **Bi**directional **K**nowledge **E**xploration (**BIKE**) for video recognition.



Background：CLIP

■ CLIP: A Web-scale Pre-trained Vision-Language Model

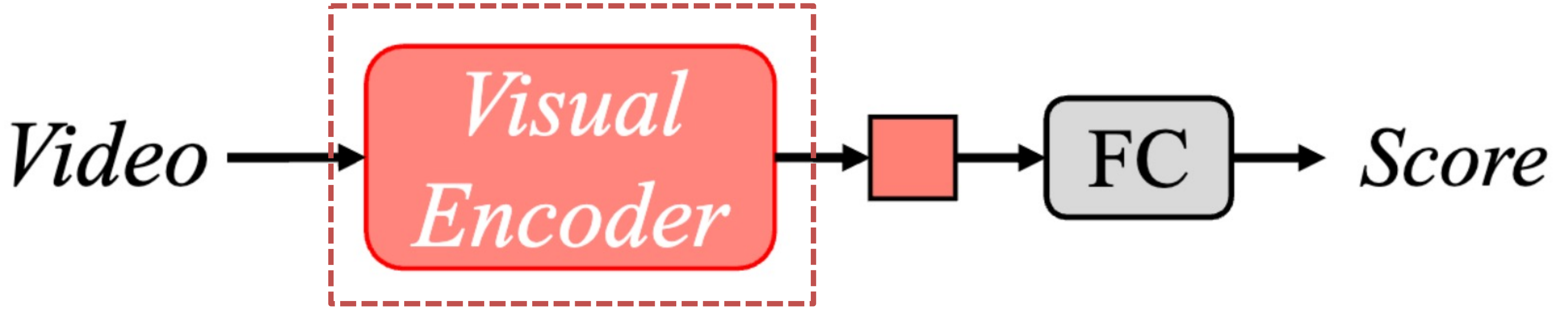


400M image-text pairs for pre-training

Radford, Alec, et al. “Learning transferable visual models from natural language supervision.” *International Conference on Machine Learning*. PMLR, 2021.

Existing Works

Vision-Only Paradigm: Traditional video recognition

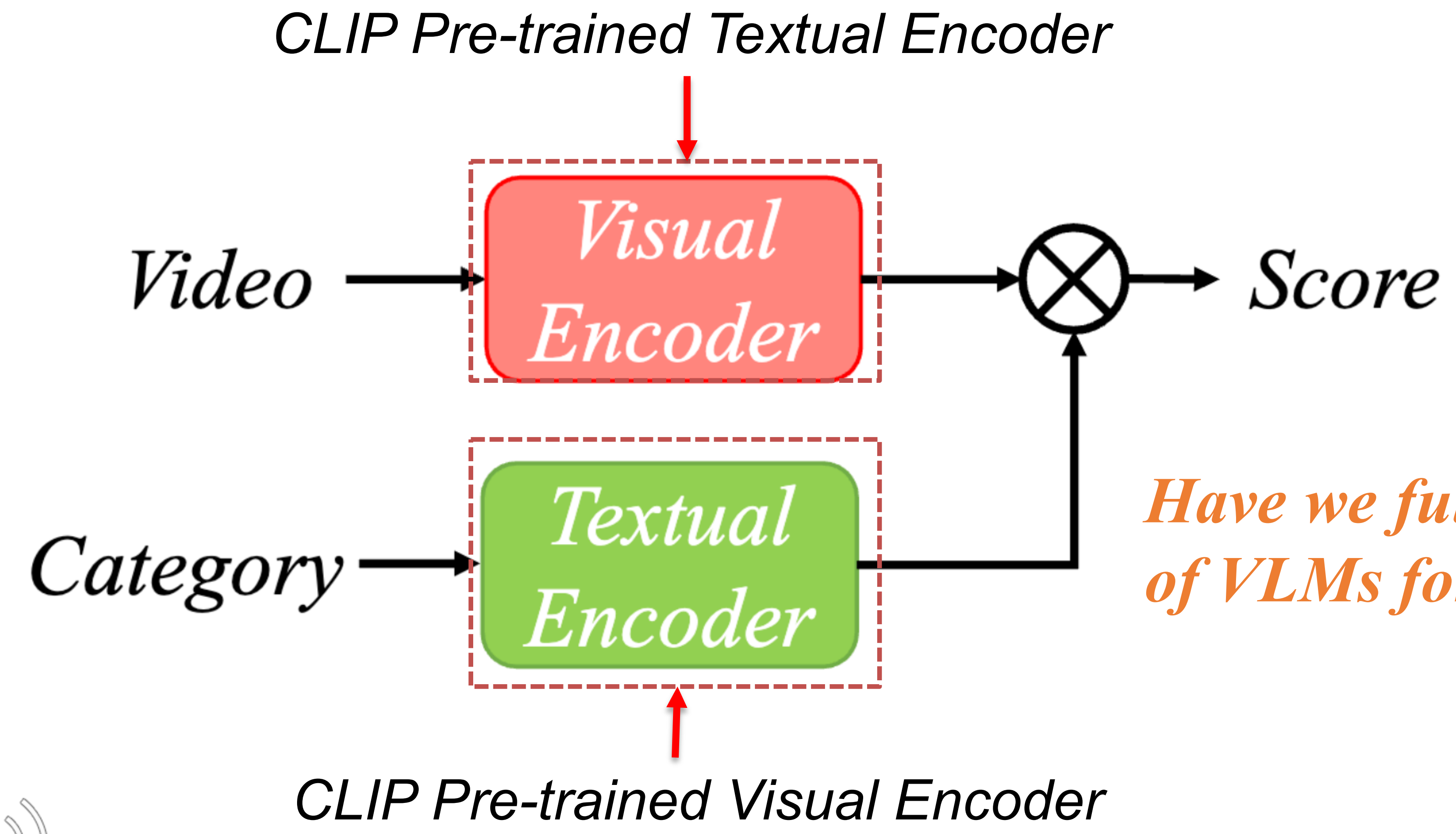


CLIP Pre-trained Visual Encoder

Limited performance on zero/few shot scenario

Existing Works

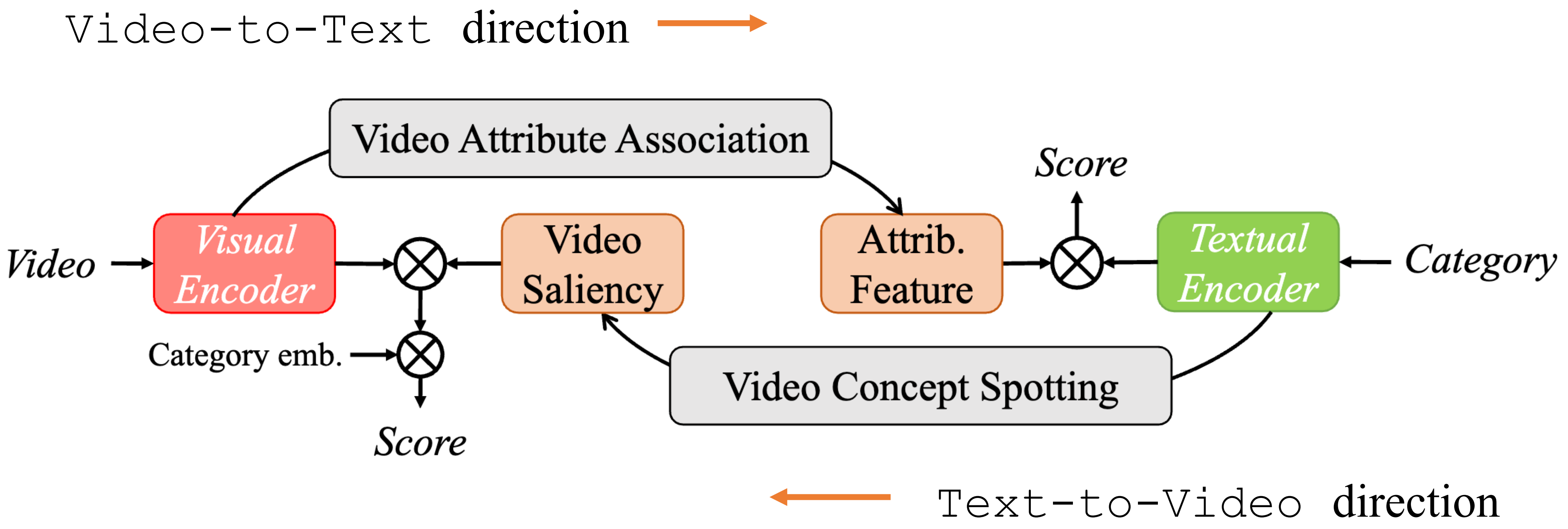
Vision-Text Paradigm: Category Embedding as Classifier



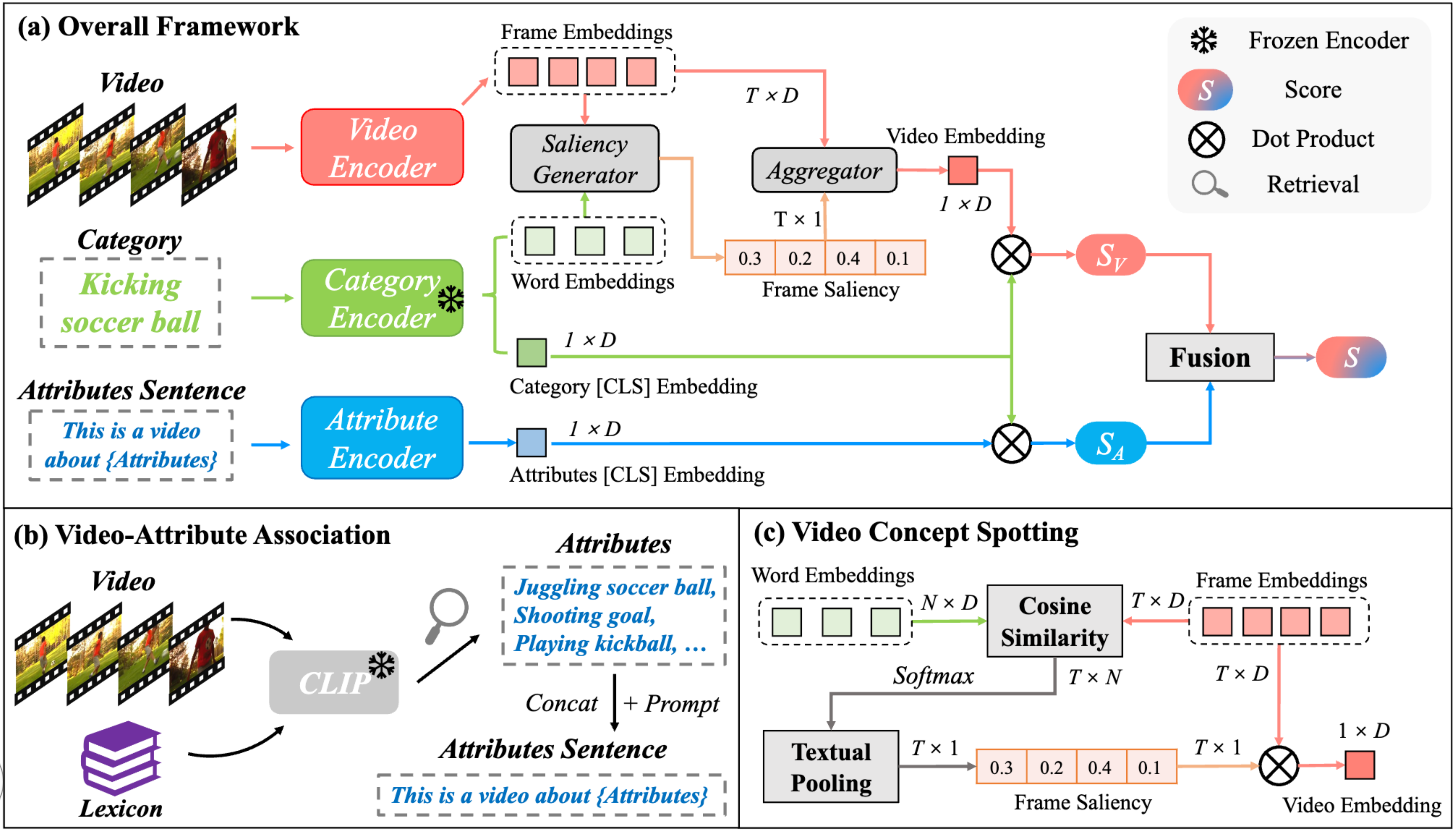
Have we fully utilized the knowledge of VLMs for video recognition?

Our BIKE

Bidirectional **K**nowledge **E**xploration (**BIKE**) for video recognition.



Our BIKE



(a) Overall Framework
Video
Video Encoder
Frame Embeddings
Saliency Generator
T × D
Aggregator
Video Embedding
1 × D
T × 1
0.3 0.2 0.4 0.1
Frame Saliency
Word Embeddings
Category
Kicking soccer ball
Category Encoder
1 × D
Category [CLS] Embedding
S_V
Fusion
S
Attributes Sentence
This is a video about {Attributes}
Attribute Encoder
1 × D
Attributes [CLS] Embedding
S_A
Frozen Encoder
Score
Dot Product
Retrieval
(b) Video-Attribute Association
Video
CLIP
Lexicon
Attributes
Juggling soccer ball, Shooting goal, Playing kickball, …
Concat + Prompt
Attributes Sentence
This is a video about {Attributes}
(c) Video Concept Spotting
Word Embeddings
N × D
Cosine Similarity
T × D
Frame Embeddings
Softmax
T × N
T × D
Textual Pooling
T × 1
0.3 0.2 0.4 0.1
Frame Saliency
T × 1
1 × D
Video Embedding

Learning Objectives

Video branch

$$\mathcal{L}_{V2C} = -\frac{1}{B}\sum_{i}^{B}\frac{1}{|\mathcal{K}(i)|}\sum_{k\in\mathcal{K}(i)}\log\frac{\exp(s(\mathbf{e_c}_i, \mathbf{e_v}_k)/\tau)}{\sum_j^B \exp(s(\mathbf{e_c}_i, \mathbf{e_v}_j)/\tau)},$$

$$\mathcal{L}_{C2V} = -\frac{1}{B}\sum_{i}^{B}\frac{1}{|\mathcal{K}(i)|}\sum_{k\in\mathcal{K}(i)}\log\frac{\exp(s(\mathbf{e_c}_k, \mathbf{e_v}_i)/\tau)}{\sum_j^B \exp(s(\mathbf{e_c}_j, \mathbf{e_v}_i)/\tau)},$$

$$\mathcal{L}_V = \frac{1}{2}(\mathcal{L}_{V2C} + \mathcal{L}_{C2V}),$$

Attributes branch

$$\mathcal{L}_{A2C} = -\frac{1}{B}\sum_{i}^{B}\frac{1}{|\mathcal{K}(i)|}\sum_{k\in\mathcal{K}(i)}\log\frac{\exp(s(\mathbf{e_c}_i, \mathbf{e_a}_k)/\tau)}{\sum_j^B \exp(s(\mathbf{e_c}_i, \mathbf{e_a}_j)/\tau)},$$

$$\mathcal{L}_{C2A} = -\frac{1}{B}\sum_{i}^{B}\frac{1}{|\mathcal{K}(i)|}\sum_{k\in\mathcal{K}(i)}\log\frac{\exp(s(\mathbf{e_c}_k, \mathbf{e_a}_i)/\tau)}{\sum_j^B \exp(s(\mathbf{e_c}_j, \mathbf{e_a}_i)/\tau)},$$

$$\mathcal{L}_A = \frac{1}{2}(\mathcal{L}_{A2C} + \mathcal{L}_{C2A}).$$

Total Loss $\mathcal{L} = \mathcal{L}_V + \mathcal{L}_A.$

Experiments

- Experimental results:
  - Comparison to the state-of-the-art methods on action recognition.
  - Comparison on multi-label video recognition.
  - Comparison on few-shot video recognition.
  - Comparison on zero-shot video recognition.

- Datasets:
  - **Kinetics-400**: ~240K videos across 400 action categories;
  - **Kinetics-600**: ~480K videos from 600 action categories;
  - **UCF-101**: 13,320 videos, 101 realistic action categories;
  - **HMDB-51**: 6,849 videos, 51 action classes.
  - **ActivityNet-v1.3**: 19,994 untrimmed videos, 200 activity categories;
  - **Charades**: 10K videos, 157 action classes.

Experimental Results

- Comparisons with SOTAs on Action Recognition

Results on ActivityNet dataset

Method	Top-1	mAP
ListenToLook [17]	-	89.9
MARL [55]	85.7	90.1
DSANet [59]	-	90.5
TSQNet [60]	88.7	93.7
NSNet [61]	90.2	94.3
BIKE ViT-L	**94.7**	**96.1**

Results on Charades dataset

Method	Frames	mAP
MultiScale TRN [71]	-	25.2
STM [20]	16	35.3
SlowFast R101 [14]	16+64	42.5
X3D-XL (312↑) [13]	16	43.4
ActionCLIP [48]	32	44.3
BIKE ViT-L	**16**	**50.4**

Results on Kinetics-400 dataset

Method	Venue	Input	Pre-training	Top-1(%)	Top-5(%)	Views	FLOPs	Param
NL I3D-101 [49]	CVPR'18	128×224^2	ImageNet-1K	77.7	93.3	10×3	359×30	61.8
MVFNet$_{En}$ [54]	AAAI'21	24×224^2	ImageNet-1K	79.1	93.8	10×3	188×30	-
TimeSformer-L [2]	ICML'21	96×224^2	ImageNet-21K	80.7	94.7	1×3	2380×3	121.4
ViViT-L/16×2 [1]	ICCV'21	32×320^2	ImageNet-21K	81.3	94.7	4×3	3992×12	310.8
VideoSwin-L [30]	CVPR'22	32×384^2	ImageNet-21K	84.9	96.7	10×5	2107×50	200.0
Methods with large-scale image pre-training								
ViViT-L/16×2 [1]	ICCV'21	32×320^2	JFT-300M	83.5	95.5	4×3	3992×12	310.8
ViViT-H/16×2 [1]	ICCV'21	32×224^2	JFT-300M	84.8	95.8	4×3	8316×12	647.5
TokenLearner-L/10 [40]	NeurIPS'21	32×224^2	JFT-300M	85.4	96.3	4×3	4076×12	450
MTV-H [63]	CVPR'22	32×224^2	JFT-300M	85.8	96.6	4×3	3706×12	-
CoVeR [68]	arXiv'21	16×448^2	JFT-300M	86.3	-	1×3	-	-
CoVeR [68]	arXiv'21	16×448^2	JFT-3B	87.2	-	1×3	-	-
Methods with large-scale image-language pre-training								
CoCa ViT-giant [65]	arXiv'22	6×288^2	JFT-3B+ALIGN-1.8B	88.9	-	-	-	2100
VideoPrompt ViT-B/16 [21]	ECCV'22	16×224^2	WIT-400M	76.9	93.5	-	-	-
ActionCLIP ViT-B/16 [48]	arXiv'21	32×224^2	WIT-400M	83.8	96.2	10×3	563×30	141.7
Florence [66]	arXiv'21	32×384^2	FLD-900M	86.5	97.3	4×3	-	647
ST-Adapter ViT-L/14 [35]	NeurIPS'22	32×224^2	WIT-400M	87.2	97.6	3×1	8248	-
AIM ViT-L/14 [64]	ICLR'23	32×224^2	WIT-400M	87.5	97.7	3×1	11208	341
EVL ViT-L/14 [27]	ECCV'22	32×224^2	WIT-400M	87.3	-	3×1	8088	-
EVL ViT-L/14 [27]	ECCV'22	32×336^2	WIT-400M	87.7	-	3×1	18196	-
X-CLIP ViT-L/14 [34]	ECCV'22	16×336^2	WIT-400M	87.7	97.4	4×3	3086×12	-
Text4Vis ViT-L/14 [58]	AAAI'23	32×336^2	WIT-400M	87.8	97.6	1×3	3829×3	230.7
		16×224^2		88.1	97.9	4×3	830×12	230
BIKE ViT-L/14	CVPR'23	8×336^2	WIT-400M	88.3	98.1	4×3	932×12	230
		16×336^2		**88.7**	**98.4**	4×3	1864×12	230

Experimental Results

- Comparisons on **few-shot** action recognition across four video datasets.

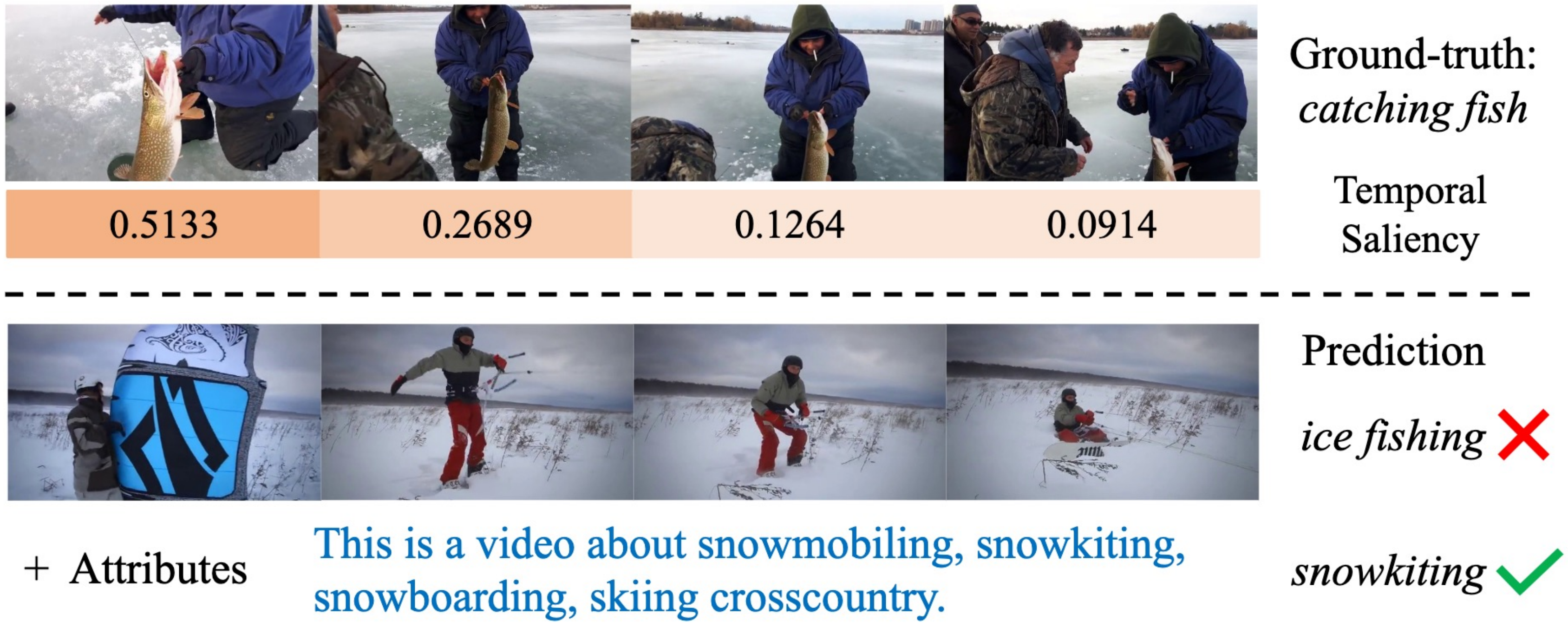
Method	**Shot**	**HMDB**	**UCF**	**ANet**	**K400**
VideoSwin [30]	2	20.9	53.3	-	-
VideoPrompt [21]	5	56.6	79.5	-	58.5
X-Florence [34]	2	51.6	84.0	-	-
	1	72.3	95.2	86.6	73.5
BIKE ViT-L	2	**73.5**	**96.1**	**88.7**	**75.7**
	5	**77.7**	**96.5**	**90.9**	**78.2**

- Comparisons on **zero-shot** video recognition across four video datasets.

Method	**UCF* / UCF**	**HMDB* / HMDB**	**ActivityNet*/ ActivityNet**	**Kinetics-600**
GA [33]	17.3±1.1 / -	19.3±2.1 / -	-	-
TS-GCN [16]	34.2±3.1 / -	23.2±3.0 / -	-	-
E2E [3]	44.1 / 35.3	29.8 / 24.8	26.6 / 20.0	-
DASZL [23]	48.9±5.8 / -	- / -	-	-
ER [8]	51.8±2.9 / -	35.3±4.6 / -	-	42.1±1.4
ResT [26]	58.7±3.3 / 46.7	41.1±3.7 / 34.4	32.5 / 26.3	-
BIKE ViT-L	**86.6±3.4 / 80.8**	**61.4±3.6 / 52.8**	**86.2±1.0 / 80.0**	**68.5±1.2**

$*$ denotes randomly selecting half of the test dataset's classes for evaluation, repeating the process ten times, and reporting the mean accuracy with standard deviation.

Visualization



Visualization of (**Top**) temporal saliency and (**Bottom**) attributes.

Conclusion

- We propose a novel framework called **BIKE** that explores bidirectional knowledge from pre-trained vision-language models for video recognition.

- In the `Video-to-Text` direction, we introduce the **Video-Attributes Association** mechanism to generate extra attributes for complementary video recognition.

- In the `Text-to-Video` direction, we introduce the **Video Concept Spotting** mechanism to generate temporal saliency, which is used to yield the compact video representation for enhanced video recognition.

- Our BIKE achieves state-of-the-art performance in most scenarios, e.g., general, zero-shot, and few-shot recognition.

THANKS

🔥 Codes & Models

https://github.com/whwu95/BIKE

👫 Contact

Wenhao Wu

Email: whwu.ucas@gmail.com

Homepage:

https://whwu95.github.io